UCSF cancer center

Magellan: Web Based Analysis Of Cancer Genomics Data

Data Analysis & Statistical Tools SIG
November 5, 2004

Chris Kingsley
ckingsley@cc.ucsf.edu
Jain Lab
UCSF Cancer Center

UCSF cancer center

Analysis of High Throughput Biological Data

Motivation

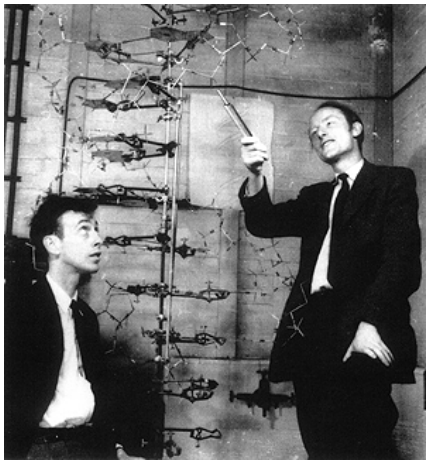
- Many researchers at the UCSF CC were moving toward high throughput methodologies in a number of different areas
  - Array based mRNA expression, CGH, proteomics, methylomics
- How are most biologists dealing with these megavariate data sets?
  - Various algorithms/applications have been appearing
    - Excel macros, SAM, Spotfire, Bioconductor, custom apps
    - Range of functionality / usability / intimidation
  - Biostatistics gurus
- How do we deliver the functionality while exploiting the domain knowledge of the biologists?

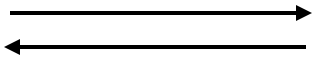
UCSF cancer center

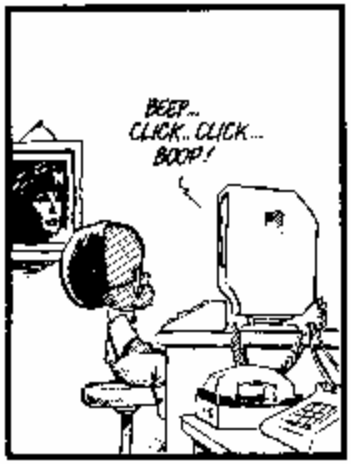
Analysis of High Throughput Biological Data

Motivation

- Many researchers at the UCSF CC were moving toward high throughput methodologies in a number of different areas
  - Array based mRNA expression, CGH, proteomics, methylomics
- How are most biologists dealing with these megavariate data sets?
  - Various algorithms/applications have been appearing
    - Excel macros, SAM, Spotfire, Bioconductor, custom apps
    - Range of functionality / usability / intimidation
  - Biostatistics gurus
- How do we deliver the functionality while exploiting the domain knowledge of the biologists?







UCSF cancer center

Build a Generalized Analytical Framework

Goals

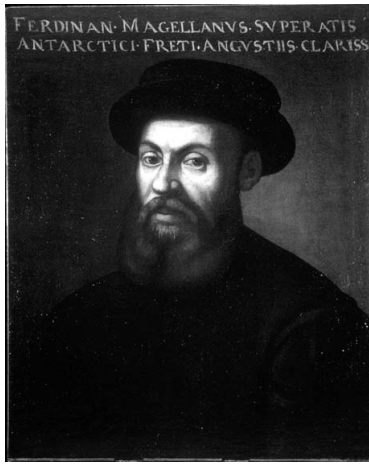
- Give biologists themselves the ability to perform analyses, such that their domain knowledge is used. Build an intuitive web based system with general application.
  - Multiple, user specified data types of arbitrary dimension
- Allow the use of biological annotation information
  - Many different quantitative / qualitative annotations can be linked to data
- Deploy analytical methods in a modular fashion for ease of extensibility
- Give users the ability to perform operations on their data prior to analysis
  - Sub selection, projection, etc.

UCSF cancer center

Build a Generalized Analytical Framework

mplementation - Magellan

- Web based
- Client – Server Model with a centralized Oracle 8 database
- Dynamic page content generated with Java/JSP
- Analytical methods generated in C or R thus far



Generality of System

Generality and Expandability is Key

- Represent data and annotations abstractly to handle as much information as possible
  - Data is derived from samples
  - Annotations describe variables of a data type
- Do not impose a nomenclature on users but insist on consistency
  - Imposing identifier nomenclature is a double edged sword – I'd rather use carrots (like databases of curated annotations) than sticks
- Do not impose particular file formats on data uploads.
- Try to minimize the pain of interfacing analytical applications
  - Provide functionality in Java Classes with a well documented API
- Provide a number of generalized operations on data that can be combined
  - Projection, sub selection, import, export, visualizations, etc.

Analytical Applications

Don’t restrict the analytical tools that can be interfaced

- Use command line access to non-Java apps, and use flat files for data transfer.
  - Other file formats can be generated by overriding Java methods
  - In the case of R, a common data structure was adopted
    - A Java method dynamically generates R code to load the flat file contents into that structure.
- Processes are forked off and the system waits for the appearance of the result file.
  - Computation done server side, but should be scalable.
  - Some results can be automatically stored as derived annotations

UCSF cancer center

Database Schema

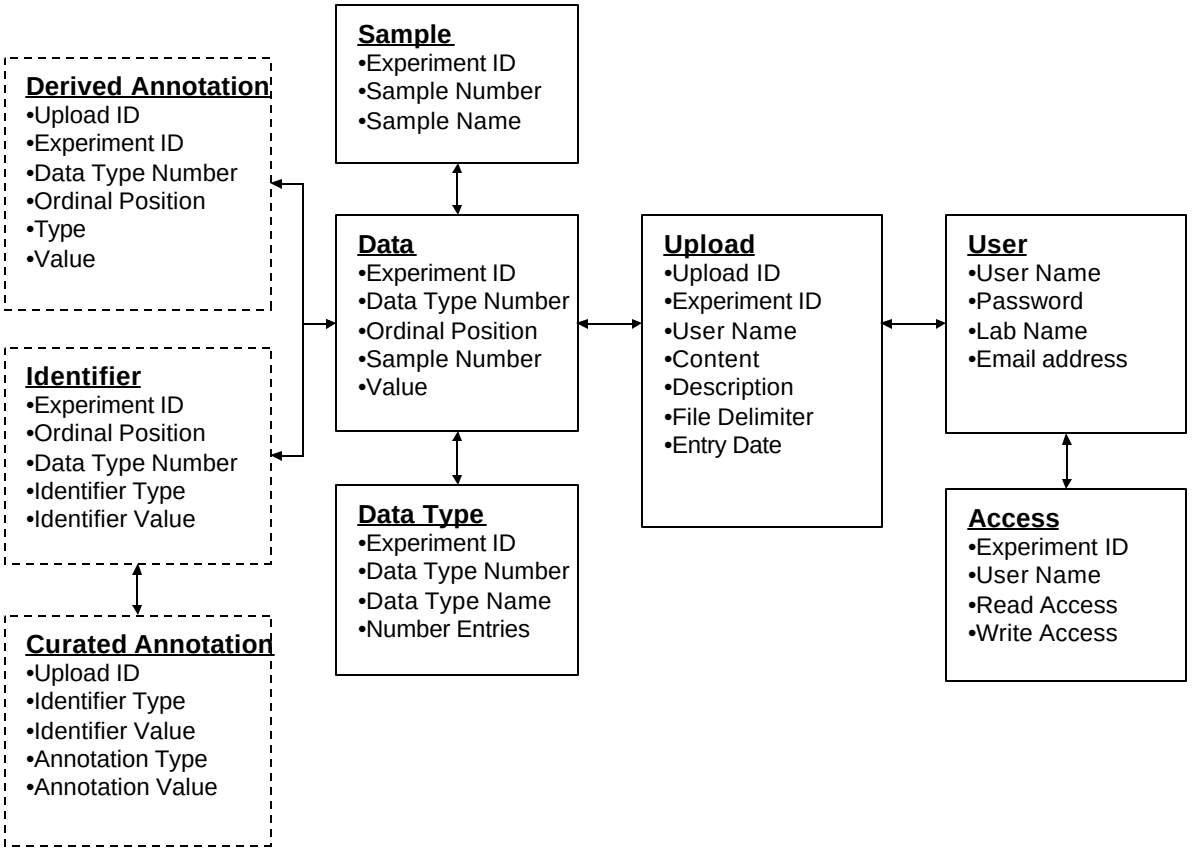
- Make no assumptions as to the type / content of data and annotations (EAV)
  - Information stored as type – value pairs of strings

UCSF cancer center

Database Schema

Make no assumptions as to the type / content of data and annotations (EAV)

- Information stored as type – value pairs of strings

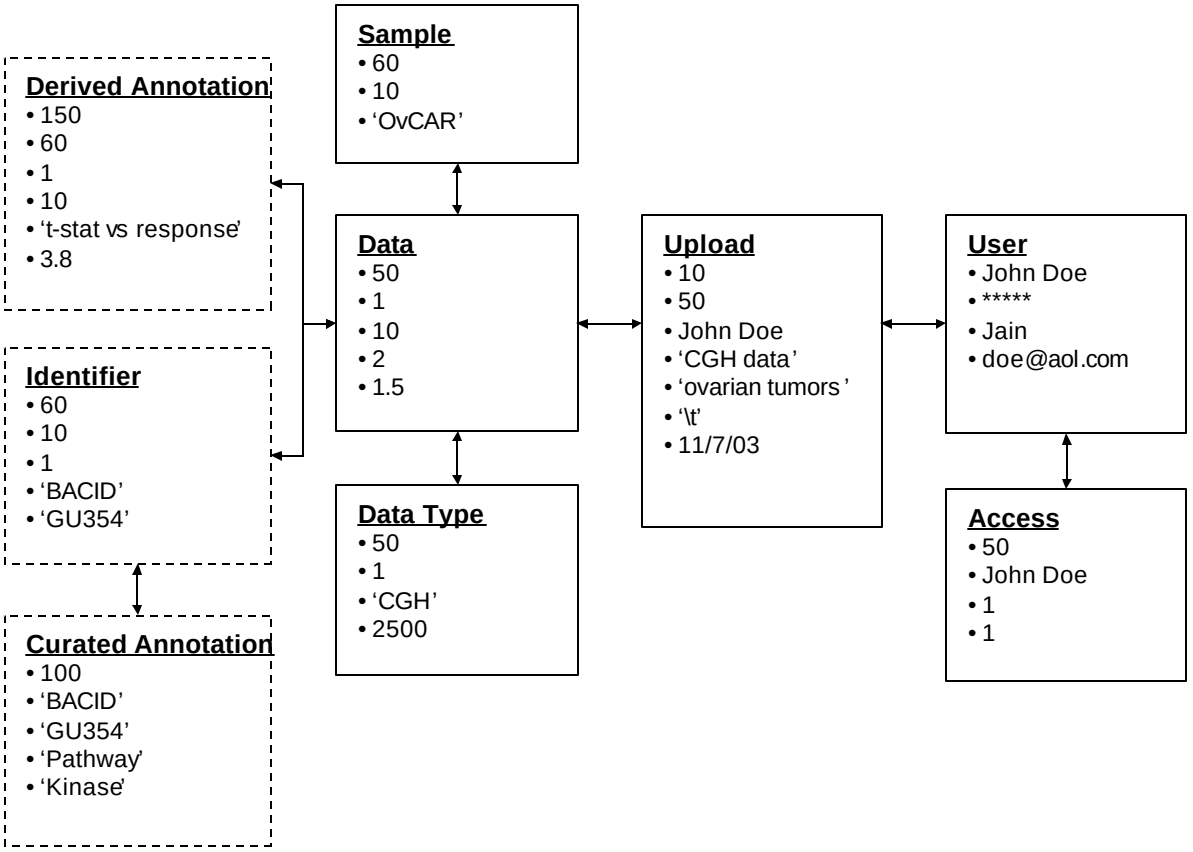


UCSF cancer center

Database Schema

Make no assumptions as to the type / content of data and annotations (EAV)

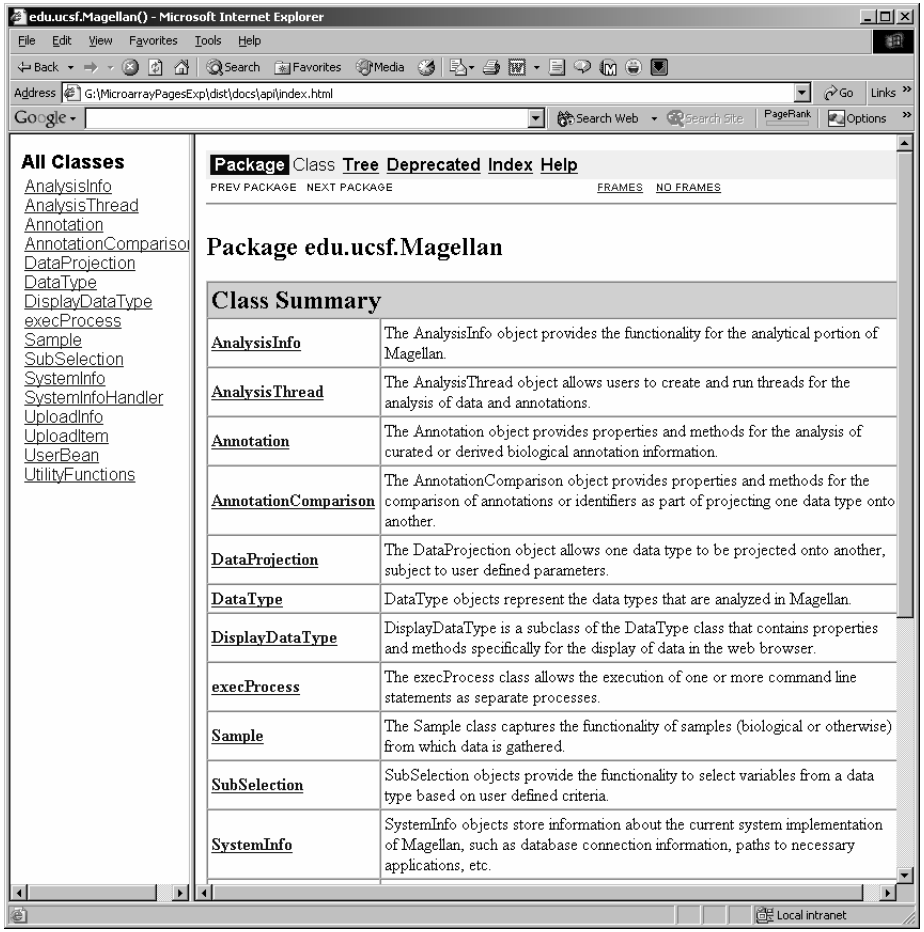
- Information stored as type – value pairs of strings



Java API

Data Representation

- All information represented by compiled Java Classes accessible from JSP pages
  - Methods allow developers to specify analytical parameters, generate data files, fork processes, etc.



UCSF cancer center

Use of Annotation Information

nnotations describe variables of a data type

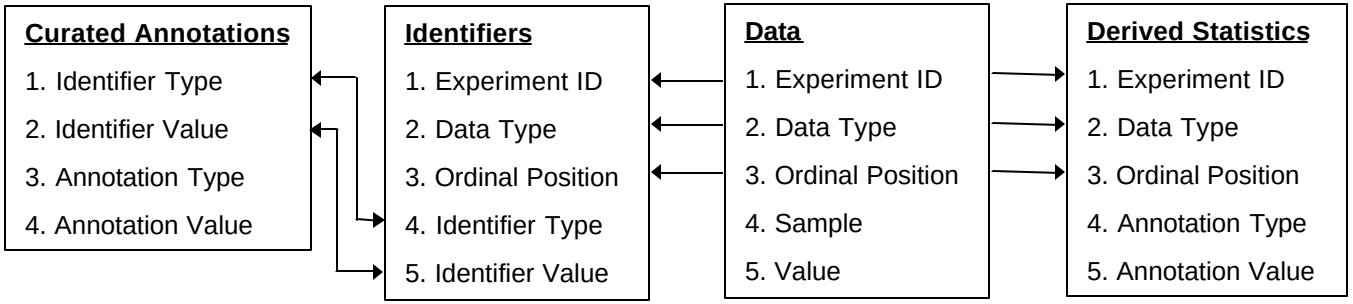
- Chromosomal position of genes, pathway designation, correlation with outcome, etc.
- Annotations can be used by certain algorithms / data operations
- Data and annotations are linked in two ways, depending on the type of annotations
  - Curated annotations – Applicable to many Data Sets. Linked through textual ‘identifiers’ such as genbank ID’s
  - Derived annotations – Specific to one data set. Linked by row number

UCSF cancer center

Use of Annotation Information

nnotations describe variables of a data type

- Chromosomal position of genes, pathway designation, correlation with outcome, etc.
- Annotations can be used by certain algorithms / data operations
- Data and annotations are linked in two ways, depending on the type of annotations
  - Curated annotations – Applicable to many Data Sets. Linked through textual 'identifiers' such as genbank ID's
  - Derived annotations – Specific to one data set. Linked by row number

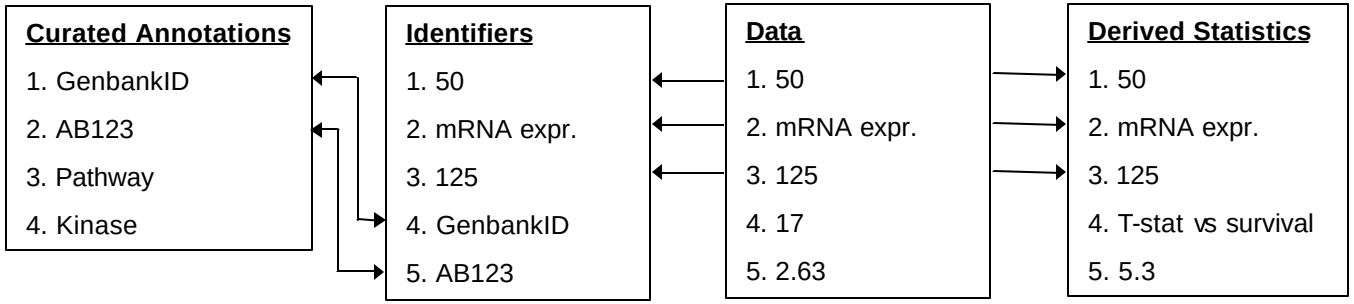


UCSF cancer center

Use of Annotation Information

nnotations describe variables of a data type

- Chromosomal position of genes, pathway designation, correlation with outcome, etc.
- Annotations can be used by certain algorithms / data operations
- Data and annotations are linked in two ways, depending on the type of annotations
  - Curated annotations – Applicable to many Data Sets. Linked through textual 'identifiers' such as genbank ID's
  - Derived annotations – Specific to one data set. Linked by row number

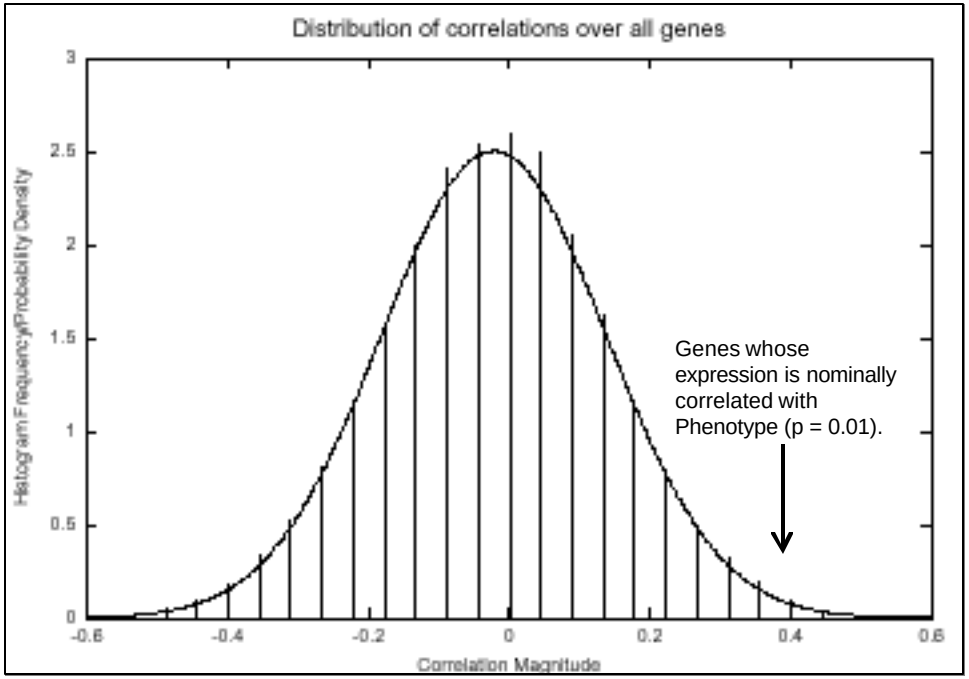


UCSF cancer center

Annotation Based Sub Selection of Data

Data Sub Selection

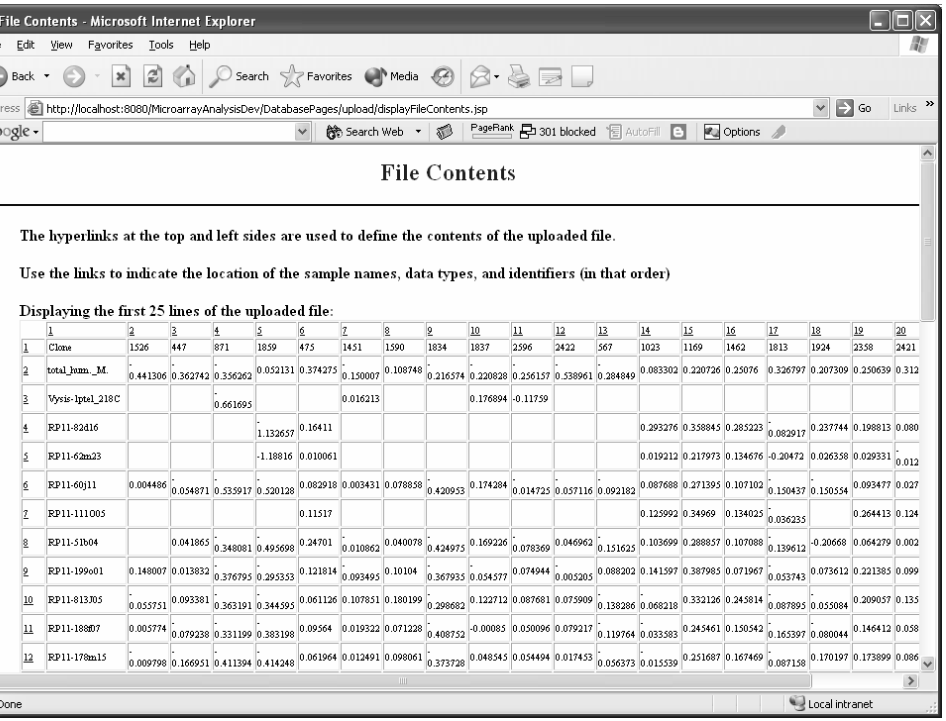
- Data sets can be sub selected based on quantitative or qualitative annotations
  - Allows the creation of biologically meaningful subsets
  - Set size reduction can reduce the effects of multiple comparisons.

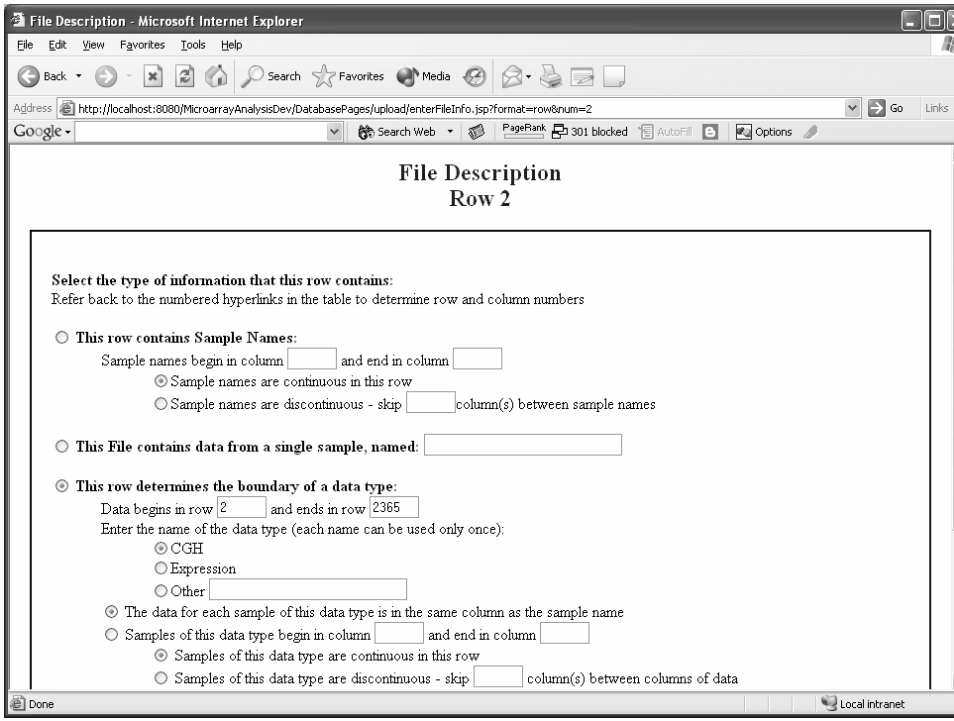


UCSF cancer center

Uploading Information

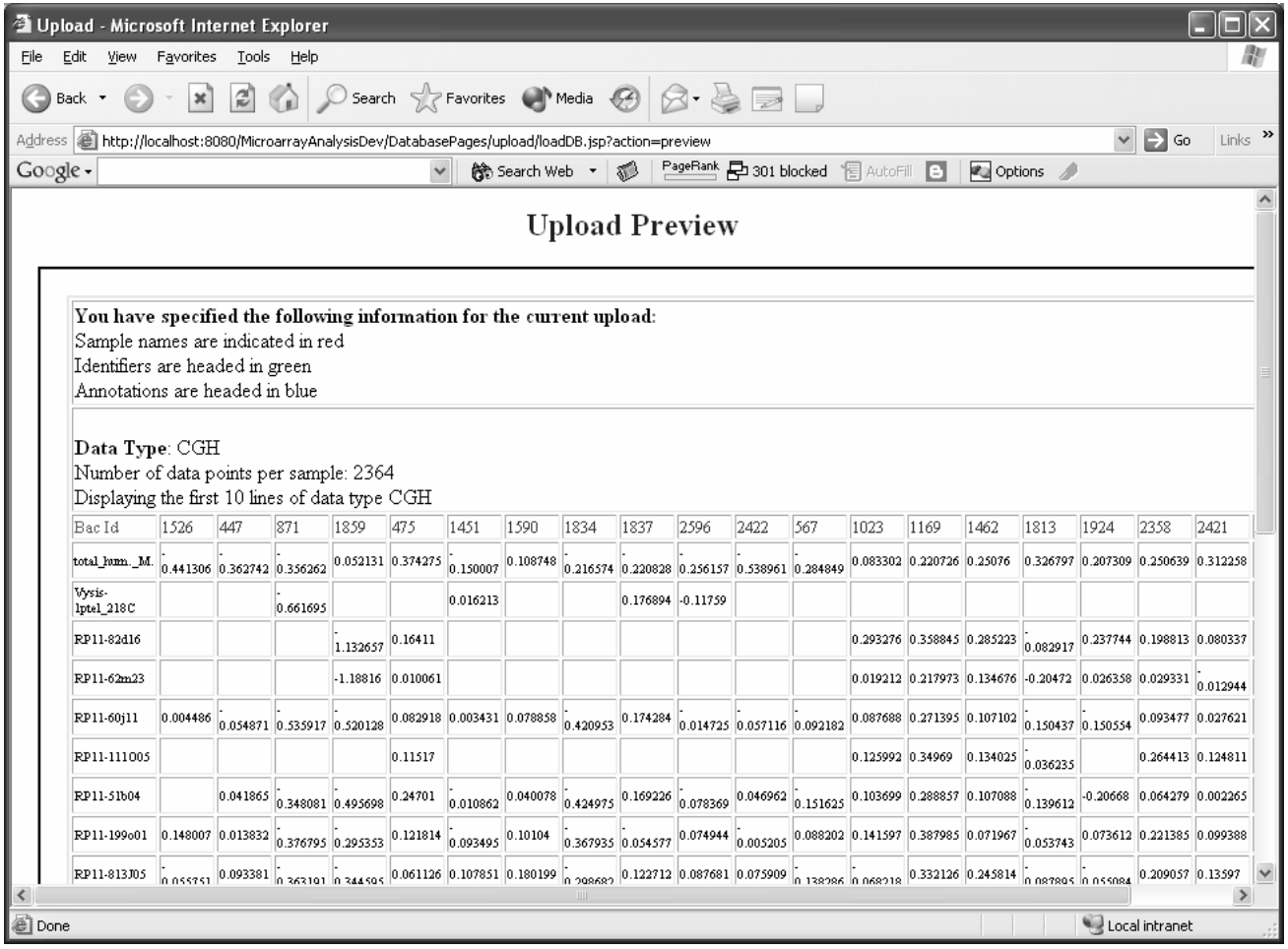
No imposed file formats - The User defines the type and location of the uploaded information





Uploading Information

Information is previewed prior to upload

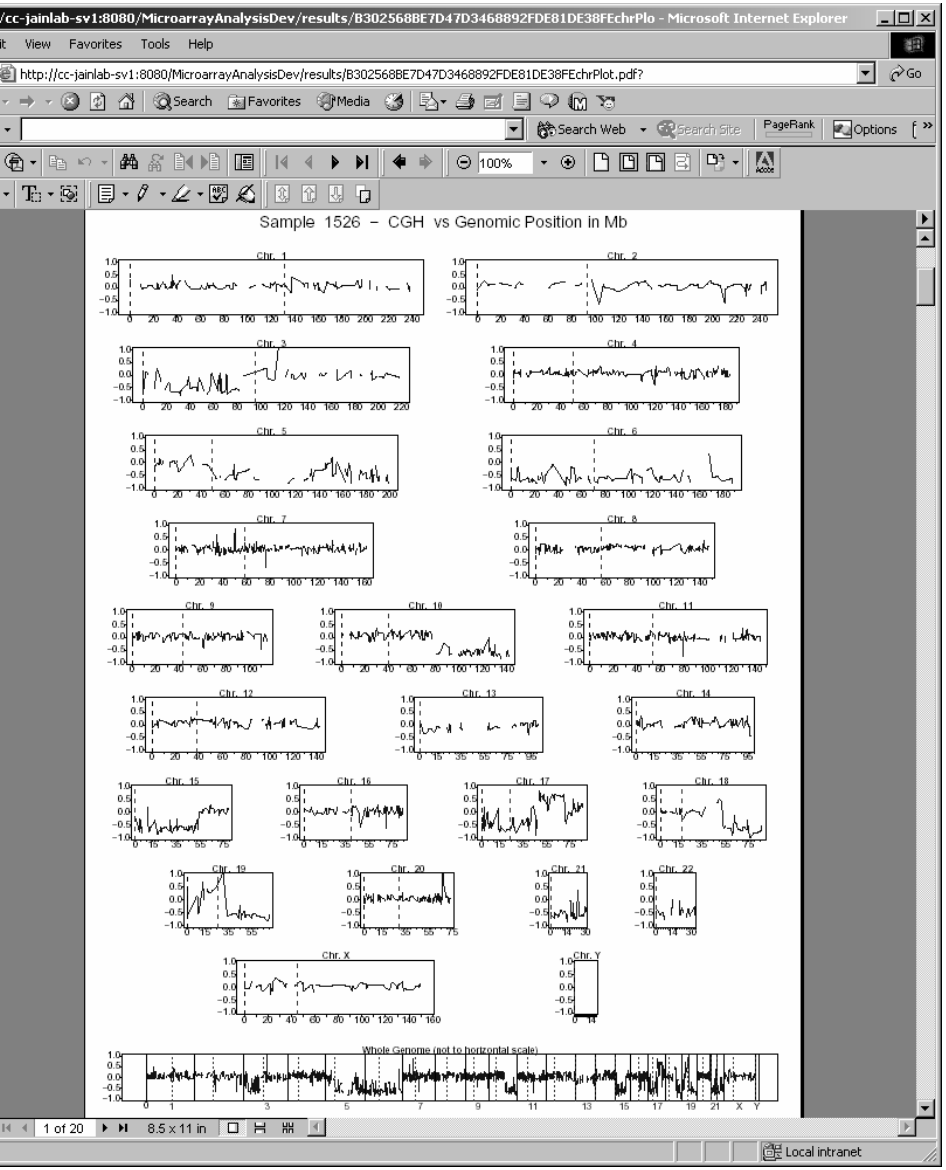


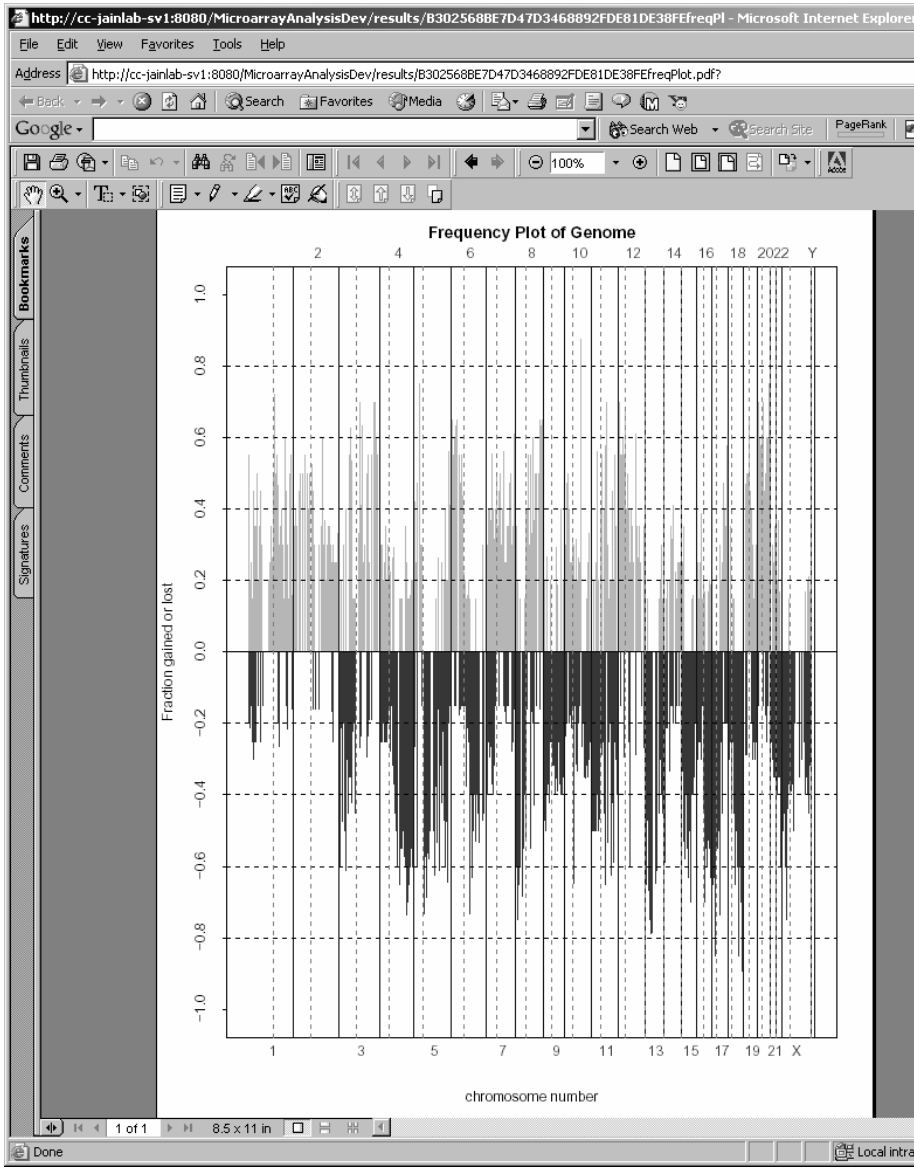
Bac Id	1526	447	871	1859	475	1451	1590	1834	1837	2596	2422	567	1023	1169	1462	1813	1924	2358	2421
total_hum._M.	-0.441306	-0.362742	-0.356262	0.052131	0.374275	-0.150007	0.108748	-0.216574	-0.220828	-0.256157	-0.538961	-0.284849	0.083302	0.220726	0.25076	0.326797	0.207309	0.250639	0.312258
Vysis-Iptel_218C			-0.661695			0.016213			0.176894	-0.11759									
RP11-82d16				-1.132657	0.16411								0.293276	0.358845	0.285223	-0.082917	0.237744	0.198813	0.080337
RP11-62m23				-1.18816	0.010061								0.019212	0.217973	0.134676	-0.20472	0.026358	0.029331	-0.012944
RP11-60j11	0.004486	-0.054871	-0.535917	-0.520128	0.082918	0.003431	0.078858	-0.420953	0.174284	-0.014725	-0.057116	-0.092182	0.087688	0.271395	0.107102	-0.150437	-0.150554	0.093477	0.027621
RP11-111O05					0.11517								0.125992	0.34969	0.134025	-0.036235		0.264413	0.124811
RP11-51b04		0.041865	-0.348081	-0.495698	0.24701	-0.010862	0.040078	-0.424975	0.169226	-0.078369	0.046962	-0.151625	0.103699	0.288857	0.107088	-0.139612	-0.20668	0.064279	0.002265
RP11-199o01	0.148007	0.013832	-0.376795	-0.295353	0.121814	-0.093495	0.10104	-0.367935	-0.054577	0.074944	-0.005205	0.088202	0.141597	0.387985	0.071967	-0.053743	0.073612	0.221385	0.099388
RP11-813J05	-0.055751	0.093381	-0.363191	-0.344595	0.061126	0.107851	0.180199	-0.298682	0.122712	0.087681	0.075909	-0.138286	-0.068218	0.332126	0.245814	-0.087895	-0.055084	0.209057	0.13597

Demo

UCSF cancer center

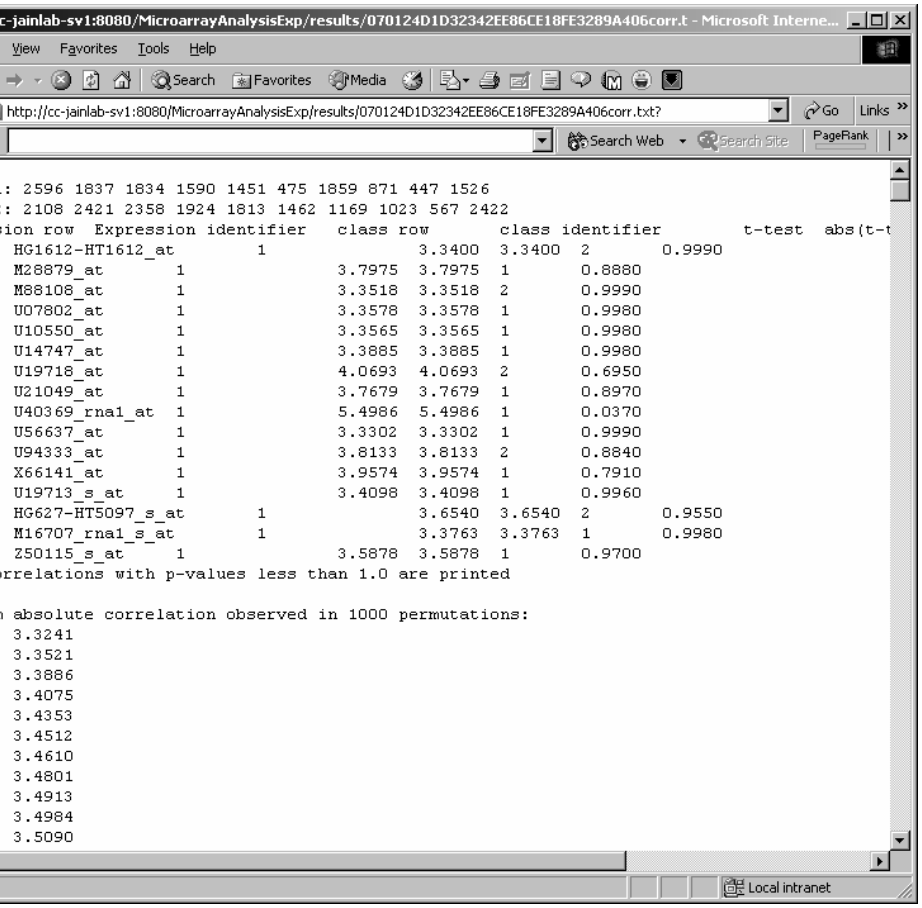
Other Analytical Functions

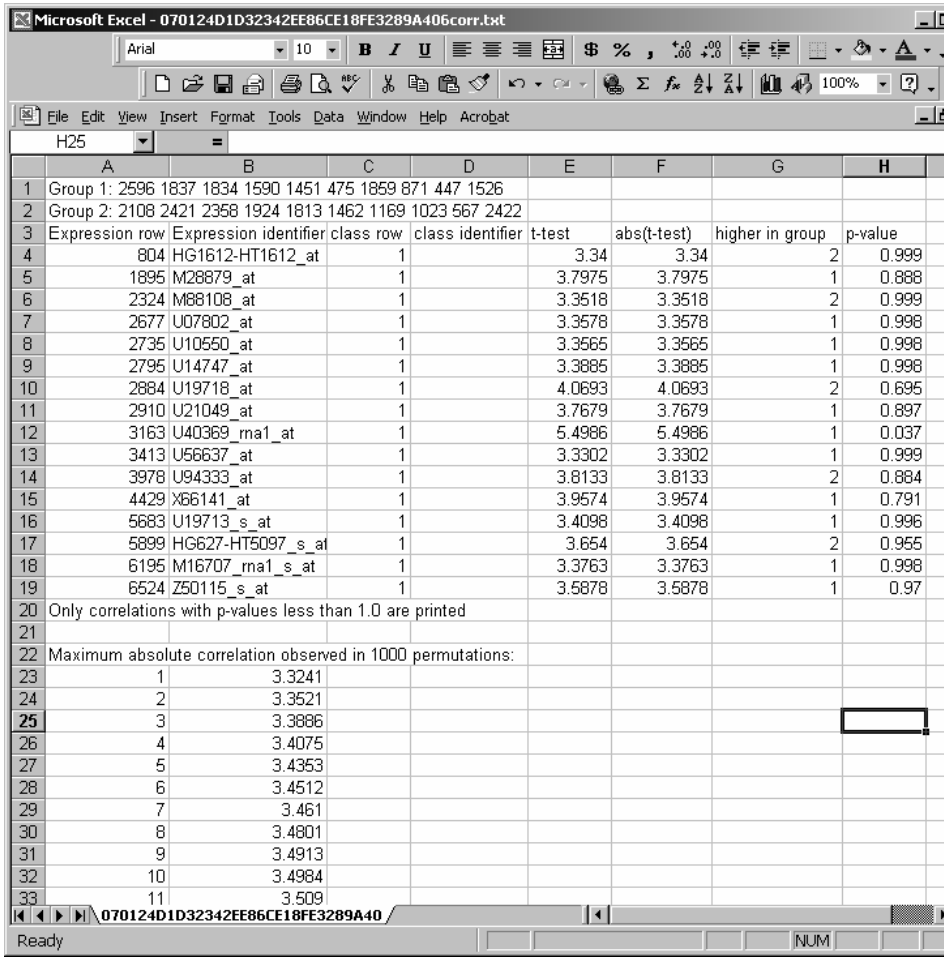




UCSF cancer center

Other Analytical Functions





UCSF cancer center

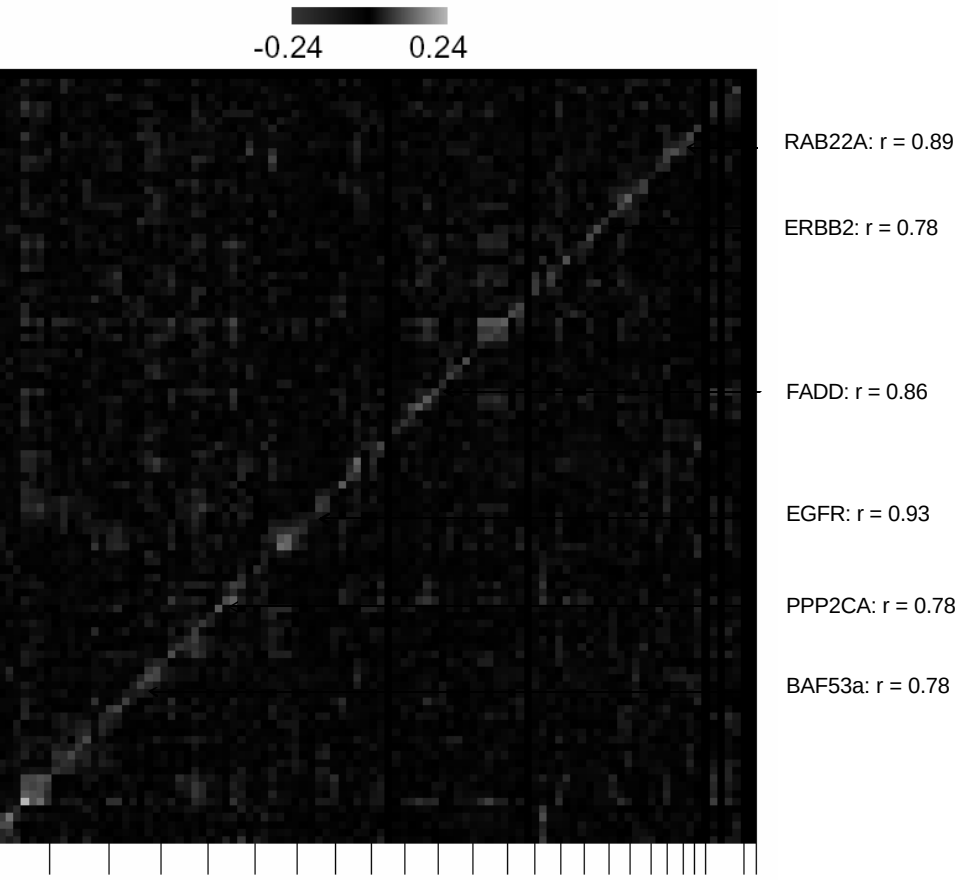
Application of Magellan to Breast Cancer Cell Line Data

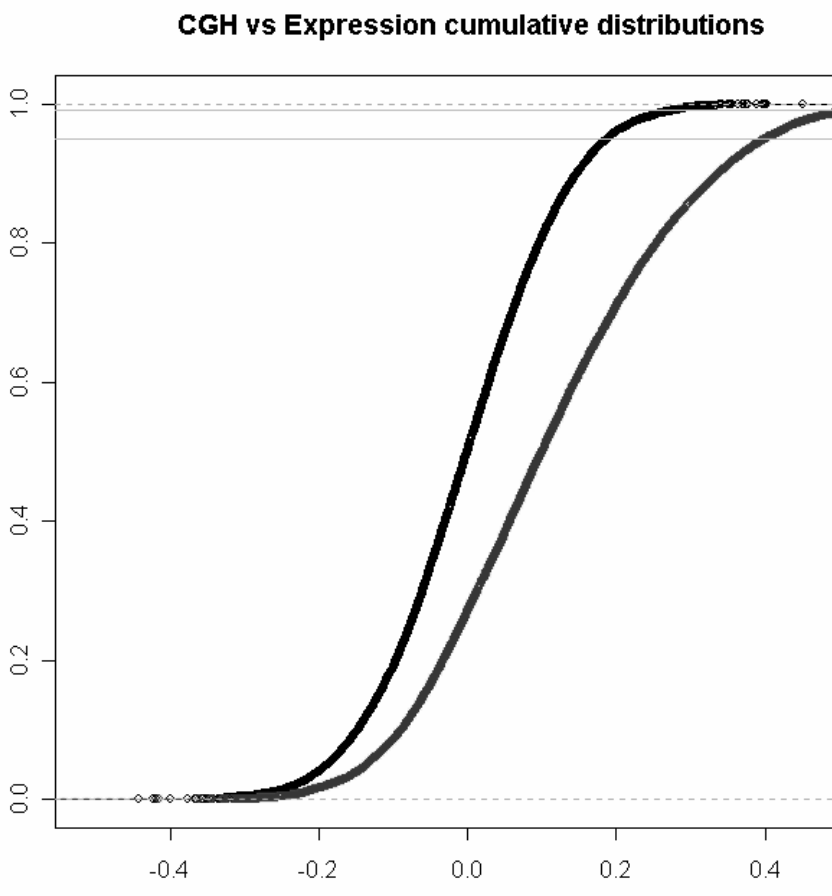
44 Breast cancer cell lines were analyzed for mRNA expression (Affy) and array based CGH

- Question: what is the effect of genomic copy number on gene expression?
- Look at sample to sample correlation of CGH/expression data, but bin by genomic position.
  - Look for genes whose expression correlates with copy number in frequently altered regions

Application of Magellan to Breast Cancer Cell Line Data

UCSF cancer center





Genome Wide Correlation Plot

- There is a positive correlation between copy number and expression. Those genes that correlate strongly can be investigated further

UCSF cancer center

Application of Magellan to Ovarian Tumor Data

Projection of subsets between data types

- Looked at CGH and mRNA expression in 20 Ovarian tumor samples (10 long, 10 short survivors)
- Used curated annotations to find ‘equivalent’ variables from one data type to another
  - Annotations can be used as a means of establishing variable equivalence
  - Equivalence is user defined (string equality, numerical comparisons, etc).

UCSF cancer center

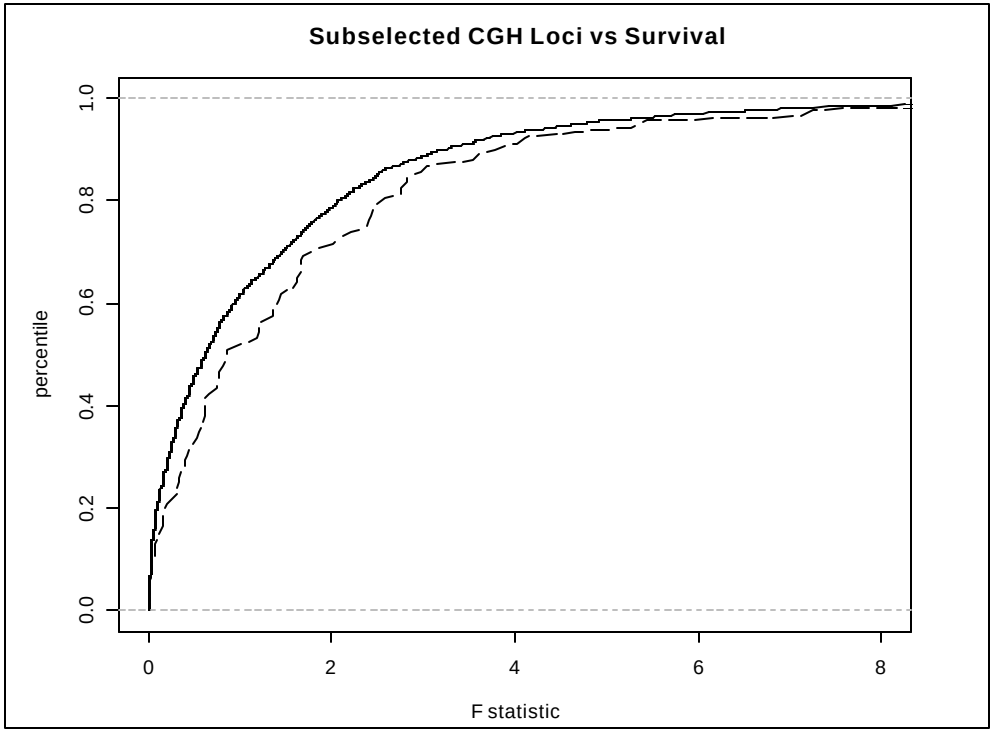
Application of Magellan to Ovarian Tumor Data

Projection of subsets between data types

- If we select for genes whose mRNA expression correlates with an outcome, do copy number changes of loci that map close to those genes also correlate?
  - Select Genes that correlate with patient survival
  - Project those genes onto CGH space – select those loci that map within 1Mb of the genes
  - Look at the correlation values of the sub selected loci vs. randomly chosen loci
- This sequence of tasks can be broken down into a series of simple operations in Magellan
  - Correlate expression with survival – store as a quantitative annotation
  - Sub select expression data
  - Project onto CGH data
  - Correlate the sub selected CGH loci with survival
  - Plot the results

UCSF cancer center

Application of Magellan to Ovarian Tumor Data



orrelation of Sub Selected Loci vs. Patient Survival

- CGH loci located in close genomic proximity to genes that correlate with survival correlate better with survival than loci chosen at random ($p<0.05$).

UCSF cancer center

Summary

Magellan allows researchers to perform visualizations and analyses of their data in a web based environment

- Abstract representation of data and annotations insures a broad applicability
- Subsetting functionality allows users to sub select data based on qualitative and quantitative annotations
  - Useful for the creation of biologically meaningful sub sets as well as a means of reducing the effects of multiple comparisons
- Analytical methods can be deployed in a modular fashion
- Generalized methods can be combined to facilitate complex analyses
  - Sub selection, projection, visualization, import, export, etc.

The Next Step

eliverables for caBIG:

- Interoperability of Magellan with caArray and caBIO
  - UML modeling of objects
  - Accessing information (especially curated annotations) from caArray
  - Decisions on use of / interface with existing caBIO objects.

ducation of End Users

- Statistics shouldn’t be a total ‘black box’ to experimentalists who are using tools like these

UCSF cancer center

Acknowledgements

Experimental collaborators

- Gray Lab
  - Daniel Pollikof, Wen-Lin Kuo
- McCormick Lab
  - Jennifer Yeh
- Andy Berchuck (Duke)

Jain Lab

- Jane Fridlyand, PhD
- Lawrence Hon
- Barbara Novak
- Adam Olshen, PhD
- Tuan Pham
- Taku Tokuyasu, PhD